Ain't Necessarily So

Coyne, James C.; Thombs, Brett D.; Hagedoorn, Mariet

Published in:
Health Psychology

DOI:
10.1037/a0017633

Health Psychology
2010, Vol. 29, No. 2, 107–116

0278-6133/10/$12.00 DOI: 10.1037/a0017633

Ain't Necessarily So: Review and Critique of Recent Meta-Analyses of Behavioral Medicine Interventions in *Health Psychology*

James C. Coyne
University of Pennsylvania School of Medicine
and University of Groningen

Brett D. Thombs
McGill University and Jewish General Hospital, Montreal, Quebec, Canada

Mariet Hagedoorn
University of Groningen

Objective: We examined four meta-analyses of behavioral interventions for adults (Dixon, Keefe, Scipio, Perri, & Abernethy, 2007; Hoffman, Papas, Chatkoff, & Kerns, 2007; Irwin, Cole, & Nicassio, 2006; and Jacobsen, Donovan, Vadaparampil, & Small, 2007) that have appeared in the Evidence Based Treatment Reviews section of *Health Psychology*. ***Design:*** Narrative review. ***Main Outcome Measures:*** We applied the following criteria to each meta-analysis: (1) whether each meta-analysis was described accurately, adequately, and transparently in the article; (2) whether there was an adequate attempt to deal with methodological quality of the original trials; (3) the extent to which the meta-analysis depended on small, underpowered studies; and (4) the extent to which the meta-analysis provided valid and useful evidence-based recommendations. ***Results:*** Across the four meta-analyses, we identified substantial problems with the transparency and completeness with which these meta-analyses were reported, as well as a dependence on small, underpowered trials of generally poor quality. ***Conclusion:*** Results of our exercise raise questions about the clinical validity and utility of the conclusions of these meta-analyses. Results should serve as a wake up call to prospective authors, reviewers, and end-users of meta-analyses now appearing in the literature.

Keywords: cancer, pain, fatigue, insomnia, arthritis

Providers of behavioral medicine services face formidable challenges justifying empirically not just *which* interventions are to be utilized, but *whether* the strength of evidence warrants that specific interventions for particular medical conditions be made available to patients and financed by third-party payments. They also need to be able to supply patients and payees with reasonable estimates of the balance of potential benefits that are likely to be obtained versus the financial and opportunity costs of behavioral treatments. Recognizing limitations of single studies and box score or narrative summaries, meta-analyses are appearing with increasing frequency in *Health Psychology* and elsewhere. These reports are widely viewed as an authoritative means to answer questions about treatment efficacy by characterizing and quantitatively synthesizing results from relevant studies, with the expectation that a representative summary effect size will result. The hope is that such meta-analyses will provide accessible, dependable, clinically informed aids to decision making.

Application of meta-analysis is facilitated by the availability of a number of user-friendly software packages. Yet, ease of use risks that results can be obtained without authors necessarily having an adequate understanding of basic principles of meta-analysis, including standards for evaluating and reporting the quality of available randomized controlled trials (RCTs) or the degree to which available RCTs are reasonably combined in a single meta-analysis. Paralleling CONSORT (Consolidated Standards of Reporting Trials; Moher, Schulz, & Altman, 2001) reporting guidelines for clinical trials, QUOROM (Quality of Reporting of Meta-analysis; Moher et al., 1999) guidelines were developed by expert consensus to improve reporting of meta-analyses and provide a checklist and flowchart for assessing the transparency with which meta-analyses are reported. Transparency, that is, presentation of basic details of how a meta-analysis was conducted, is essential so that readers can form their own independent opinion of the adequacy of methods, results, and conclusions. Over 1,100 citations of the original QUOROM paper in less than a decade indicate wide dissemination of these standards. However, not one citation of QUOROM can be found in *Health Psychology*, and there have been only two passing references in *Annals of Behavioral Medicine*, neither in the context of a meta-analysis. None of the more than 30 post-QUOROM

James C. Coyne, Department of Psychiatry, University of Pennsylvania School of Medicine, and Department of Health Sciences, Graduate School for Health Research, University Medical Center Groningen, University of Groningen; Brett D. Thombs, Department of Psychiatry, McGill University and Jewish General Hospital, Montreal, Quebec, Canada; and Mariet Hagedoorn, Department of Health Sciences, Graduate School for Health Research, University Medical Center Groningen, University of Groningen.

Dr. Thombs is supported by a New Investigator Award from the Canadian Institutes of Health Research and an Établissement de Jeunes Chercheurs award from the Fonds de la Recherche en Santé Québec.

Correspondence concerning this article should be addressed to James C. Coyne, Ph.D., Department of Psychiatry, University of Pennsylvania School of Medicine, 3535 Market St., Room 676, Philadelphia, PA 19104. E-mail: jcoyne@mail.med.upenn.edu

meta-analyses published in *Health Psychology* or *Annals of Behavioral Medicine* mentions QUOROM. This raises the question of whether authors, reviewers, or intended consumers of meta-analyses of behavioral medicine have adequate knowledge of standards for conducting and reporting meta-analyses, recent conceptual and statistical developments, or current controversies in meta-analysis methodology. It also suggests that early warnings about the perils of conducting meta-analysis as if it were a straightforward, even mechanical procedure with a minimal amount of subjectivity or arbitrary judgment by authors (Wanous, Sullivan, & Malinak, 1989) may not have been heeded.

Coyne, Thombs, and Hagedoorn (2008) recently provided a extensive critique of a meta-analysis of interventions for distress among breast cancer patients (Zimmermann, Heinrichs, & Baucom, 2007). The first research question addressed by Zimmermann et al. (2007) was whether patients with breast cancer had better outcomes when they received interventions as part of a study that only included those with breast cancer as compared with studies that included patients with mixed diagnoses. Zimmermann et al. concluded from their meta-analysis that studies that mixed the type of cancer had significantly larger effect sizes than studies limited to breast cancer. The implication would seem to be that patients with breast cancer should be provided with nonspecific psychological interventions designed for patients with any type of cancer, rather than having interventions tailored to the breast cancer experience. However, going back to the original studies, we were unable to identify a single RCT that provided data to compare results between patients with breast cancer being treated with treatment protocols designed specifically for breast cancer versus being treated with protocols designed generally for patients with cancer. In addition, no allowance was made for the poor methodological quality of studies entered into the meta-analysis. Previous authors had either abandoned efforts to conduct meta-analyses because of the poor quality of available trials (Newell, Sanson-Fisher, & Savolainen, 2002), or had substantially lowered estimates of efficacy when studies that were of poor quality or that used overly small sample sizes were excluded (Sheard & Maguire, 1999). We were unable to find a number of basic details in the article as to how the meta-analysis had been conducted. For instance, studies with multiple outcomes were assigned single effect sizes without indication of how this was done. We requested more information from the authors. Comparing these materials to the original studies, we found a substantial error in calculating an effect size, several misclassifications of interventions, and the same intervention trials being counted two and three times. Our review was not comprehensive, but was sufficient to raise doubts as to whether clinical and policy decisions should be based on the article's conclusions. We were also left wondering how pervasive the problems we observed in Zimmermann et al. were in meta-analyses of behavioral medicine interventions.

In the present paper, we undertook a similar exercise with four meta-analyses of behavioral interventions for adults (Dixon, Keefe, Scipio, Perri, & Abernethy, 2007; Hoffman, Papas, Chatkoff, & Kerns, 2007; Irwin, Cole, & Nicassio, 2006; and Jacobsen, Donovan, Vadaparampil, & Small, 2007) that appeared in a new section of *Health Psychology*, Evidence Based Treatment Reviews. Since our objective was to identify the degree to which meta-analyses of behavioral interventions are based on flawed methodology, and not to evaluate the pooled effect sizes reported in each meta-analysis per se, we did not redo these meta-analyses, but provide an evaluation of the adequacy of their conduct, reporting, and clinical recommendations. Presumably readers would face a similar task in evaluating these articles, although they might not be motivated to look beyond the meta-analyses themselves due to confidence in the objective, straightforward nature of the tasks of conducting a meta-analysis, reporting findings, and making recommendations. We started with the articles themselves and any supplementary materials, asked questions of the authors, then went to the original RCTs and related articles. We asked:

(1) Was the conduct of the meta-analysis accurately, adequately, and transparently described in the article or supplementary materials?

(2) Was there an adequate attempt to deal with the methodological quality of the original intervention trials?

(3) To what extent did the results of the meta-analysis depend on small, underpowered studies?

(4) To what extent did the meta-analysis provide valid and useful evidence-based recommendations to clinicians and policymakers?

Items 2, 3 and 4 warrant brief elaboration. The Cochrane Collaboration has spelled out guidelines for conducting meta-analyses that emphasize the necessity of taking methodological quality into account (Higgins & Altman, 2008). Serious deficiencies in RCTs of behavioral medicine interventions have been noted (Coyne, Lepore, & Palmer, 2006; Newell et al., 2002; Sheard & Maguire, 1999). While dozens of scales show validity for evaluating methodological quality, they nonetheless differ in their ratings in an arbitrary fashion (Juni, Witschi, Bloch, & Egger, 1999). There is increasing acceptance of the idea that rather than weighting trials for their methodological quality as scored with these scales, some threshold should be set as a basis for outright exclusion of trials that either do not achieve this minimal quality or for which analyses should be conducted both with and without their inclusion (Higgins & Altman, 2008; Juni, Altman, & Egger, 2001).

Our third criterion is controversial, but we believe it is important. Kraemer, Gardner, Brooks, and Yesavage (1998) propose excluding small, underpowered studies from meta-analyses. The risk of including studies with inadequate sample size is not limited to clinical and pragmatic decisions being made on the basis of trials that cannot demonstrate effectiveness when it is indeed present. Rather, Kraemer et al. demonstrate that inclusion of small, underpowered trials in meta-analyses produces gross overestimates of effect size due to substantial, but unquantifiable confirmatory publication bias from nonrepresentative small trials. Without being able to estimate the size or extent of such biases, it is impossible to control for them. Other authorities voice support for including small trials, but generally limit their argument to trials that are otherwise methodologically adequate (Sackett & Cook, 1993; Schulz & Grimes, 2005). Small trials are particularly susceptible to common methodological problems, however, such as lack of baseline equivalence of groups; undue influence of outliers on results; selective attrition and lack of intent-to-treat analyses; investigators being unblinded to patient allotment; and not having

a predetermined stopping point so investigators are able to stop a trial when a significant effect is present.

Staines and Cleland (2007) recently analyzed the psychotherapy literature and concluded that the failure to reduce the weight of studies with small sample sizes resulted in only a "minor" upward bias in effect size estimates of approximately 1.4 times unweighted effect estimates. However, the inclusion of one or several studies with small sample sizes is one thing. It is an altogether different problem when most or all of the studies included in a given meta-analysis have small sample sizes. Small trials that are not statistically significant are not usually published because of the criticism that significant effects could not have been expected, whereas a trial that obtains statistically significant effects despite being underpowered is considered noteworthy and results are readily publishable. Small trials that are published generally need to have sizable effect sizes just to meet the minimum threshold for statistical significance (e.g., $p < .05$). A trial with 20 participants in the treatment and control groups, for instance, would require an effect size of at least 0.66 for statistical significance. Even with $n = 50$ per cell a minimum effect size of 0.40 would be needed. This does not accurately describe the extent of the problem, however, since small studies that cross the $p < .05$ threshold do so by varying degrees, with some producing quite large effect sizes, even when the null hypothesis of no treatment effect is true. Kraemer et al. (1998) showed that when the true effect of a treatment is zero and $n = 20$ per subgroup, the estimated standardized mean difference effect size in a meta-analysis of statistically significant RCTs will be between 0.90 and 1.00. With $n = 50$ per subgroup and a true null finding, the expected effect size would be approximately 0.60. If small studies are only published when their results are significant, the effect sizes available in the published literature will of necessity appear moderate to large, because such effect sizes are the only ones that can reach statistical significance with small sample size. Other authors have provided similar critiques of the ramifications of combining sets of small studies to produce a single effect estimate (e.g., Ioannidis, 2008a; Howard et al., 2009).

How much power is sufficient for inclusion of a study in a meta-analysis? Kraemer et al. (1998) suggest at least 50% power, but note that proposals for individual studies usually require at least 70% power. Kraemer (personal communication, December 8. 2008) proposes that trials have at least a .70 probability of detecting a moderate size effect (e.g., $\delta = .50$ based on Cohen, 1992), requiring at least 50 patients per group. Unfortunately, a large proportion of studies of behavioral medicine interventions do not meet this criterion. For instance, only 16 of 56 (29%) studies reviewed by Zimmermann and colleagues (2007), included at least 50 patients per cell. Thus, we reluctantly will apply a more liberal criterion of 35 patients per cell, with 55% power of detecting a moderate effect.

Our fourth criterion involves evaluating the pragmatic significance of findings of a meta-analysis. This criterion takes into account our other three criteria, but also the clinical heterogeneity of the patients, outcomes, interventions, and study procedures that contributed to the overall effect size. The degree to which a summary effect size represents a set of trials with similar study characteristics reflects how well one might expect the overall effect size to generalize to any given intervention. Ioannidis (2008b) argues cogently that a lack of statistical heterogeneity often does not indicate a lack of clinical heterogeneity. In our review of Zimmermann et al.'s meta-analysis (2007), for instance, we found that composite effect sizes were constructed from, among other things, scores on a measure of patient satisfaction with a cancer center tour (McQuellon et al., 1998), the immediate mood of patients with breast cancer following a cosmetics class with free cosmetics (Manne, Girasek, & Ambrosino, 1994), and Hamilton Depression Rating Scale scores for patients who received a problem-solving intervention after having been prescreened for high distress with a threshold that excluded most patients (Nezu, Nezu, Felgoise, McClure, & Houts, 2003). Thus, we ask to what degree clinically meaningful generalizations can be made from the summary effect estimate back to the patients, from interventions and outcomes of each of the individual studies that were synthesized to generate the summary measure.

In the following sections, we review each of the four meta-analyses in question based on this set of criteria, beginning with the meta-analysis by Irwin et al. (2006) and followed by the meta-analyses by Hoffman et al. (2007); Dixon et al. (2007), and Jacobsen et al. (2007). In each section, we provide an overview of the meta-analysis, followed by a description of problems identified in the conduct and interpretation of the meta-analysis, then a critique of the authors' conclusions and a message to consumers of the meta-analysis.

Irwin et al. (2006)

Overview

Irwin and colleagues conducted a meta-analysis of behavioral interventions for insomnia with a stated objective of "comparing responses in studies that exclusively enrolled persons who were 55 years of age or older versus outcomes in randomized controlled trials that enrolled adults who were, on average, younger than 55 years of age" (p. 4). They retrieved 51 articles, of which 23 were included in the meta-analysis. The findings revealed significant overall effects of behavioral interventions for sleep quality, latency, efficiency, and wakening after sleep onset ($d = 0.50$ to 0.79), and a nonstatistically significant effect for total sleep time ($d = 0.17$). Some moderator effects were found for intervention type and age.

Transparency of Reporting

Irwin et al.'s review of behavioral interventions for primary insomnia was published before *Health Psychology's* webpage was available for providing supplementary material, and presentation of basic details of the included RCTs was likely compromised by limitations on article length. Nonetheless, details important for the evaluation and interpretation of the results of the meta-analysis were not available for readers. For instance, the authors indicated that decisions whether to include studies were based on methodological quality, but no description of the evaluation and decision process was provided.

Sample Sizes of Original Studies and Other Problems

Exclusion of small trials ($n < 35$) would have eliminated *all* eight studies of older adults; five of these studies included 15 or

fewer participants per condition. Of the studies including younger adults, 14 of the 15 studies had fewer than 35 participants per condition. Furthermore, most moderator analyses comparing older versus younger patients and different intervention types were based on only a few studies per subgroup (median = 5). Even if one accepts the small sample sizes of the original studies, the significant moderator effects are disputable. For example, the analysis purported to show an effect for intervention type on sleep efficiency compared three cognitive–behavioral therapy (CBT) studies (ES = 1.47, 95% CI = 1.00–1.94) to two relaxation studies (ES = −0.35, 95% CI = −0.75–0.05) of older adults. The effect size (d = 2.64) for the Rybarczyk, Lopez, Benson, Alsten, and Stepanski (2002) CBT study seems to be exceptionally high. Also, Irwin et al. appear to have included the effect size of the Sleep Compression group (d = −0.02) instead of the Relaxation group (d = 0.16) of the study by Lichstein Riedel, Wilson, Lester, and Aguillard (2001).

Message to Consumers

Irwin et al. (2006) offered an optimistic assessment: "The review supported the efficacy of behavioral interventions . . . The magnitudes of the effect sizes were substantial" and "CBT proved to be substantially more effective than relaxation training in improving sleep efficiency" (p. 10). Further: ". . . the current meta-analyses confirmed the general efficacy of behavioral interventions across cohorts with two exceptions. Behavioral interventions were more effective in the younger cohort in TST and efficiency in the older cohort" (p. 11). These claims, however, are based on a collection of studies that include only one study with 35 or more participants per cell. Even in the case of a true null, such a collection of published small studies would be expected to produce a robust effect size estimate (Howard et al., 2009; Ioannidis, 2008a; Kraemer et al., 1998). Furthermore, appraisal of these claims requires information unavailable in the article that only a skeptical reader would likely seek.

Hoffman et al. (2007)

Overview

In a meta-analysis of psychological interventions for chronic low back pain, Hoffman et al. (2007) reported that "a total of 205 effect sizes from 22 studies were pooled in 34 analyses" and that "positive effects of interventions were found for pain intensity, pain-related interference, health-related quality of life, and depression" (pp. 1). Hoffman and colleagues retrieved 96 articles, of which 39 met inclusion criteria, and 34 (representing 31 studies) had extractable data for the meta-analysis. Another nine studies were eliminated, mostly because they featured contrasts between interventions considered by Hoffman et al. for the purposes of their meta-analysis to be identical.

Transparency of Reporting

Overall, data from 22 studies were entered into the meta-analysis. However, tracking which studies were entered into particular comparisons across the various tables in the article became a frustrating task because numbers did not match up. In some cases, this could have occurred because data were not available for particular outcomes, but in other instances, this could not have been the explanation. To cite one of a number of examples, for post-treatment pain interference, the number of comparisons between intervention and active control (k) and number of patients (N) are the same in Table 2 and 3, but d = −.10 in Table 2 and .20 in Table 3.

Methodological Quality of Included Studies

Focusing on material presented in the article alone, we found a lack of evidence that psychological interventions were superior to other active treatments and no evidence of enduring effects of psychological interventions beyond immediate posttreatment assessments. Moreover, based on a table of methodological ratings that the authors graciously provided, we found only five of 22 RCTs met more than half of the 12 applicable methodological criteria that were applied by the authors. The largest RCT (Alaranta et al., 1994) met 17%. Further examining the table provided by the authors, we found that 60% of the studies involved intervention and control groups that were not comparable on key variables at baseline; less than half of the studies adequately indicated the number of patients enrolled, treatment drop-out and reasons for drop-outs; and only 15% of trials provided intent-to-treat analyses. Less than a third had manualized treatment procedures or detailed protocols; only three assessed patient adherence to prescribed activities and only three restricted outside interventions.

Sample Sizes of Original Studies and Other Problems

Turning to the original studies, we found that 17 of the 22 studies included by Hoffman fell below $n \geq 35$ per group. Exclusively psychological interventions were distinguished from multimodal interventions in which a psychological intervention was embedded, but no study of a multicomponent intervention allowed evaluation of the independent contribution of a psychological component. This cointervention confound (Cochrane Collaboration; O'Connor, Green, & Higgins, 2008) left no way of distinguishing if psychological components were superfluous, additive, or crucially decisive. Moreover, there was often no indication that the multicomponent treatments applied psychological principles or required psychologically trained interventionists. The details provided were often sparse, but one trial described the psychological component as "a lecture to give the patient an understanding that ordinary physical activity would not harm the disk and a recommendation to use the back and bend it" (Brox et al., 2003). Of the studies providing effect sizes for a comparison between a psychological intervention and an active control treatment, three were small, underpowered studies (Hernandez-Reif, Field, Krasnegor, & Theakston, 2001, n = 12 for each group; Kankaanpaa, Taimela, Airaksinen, & Hanninen, 1999, n = 30 and 24 for the experimental and control group, respectively; Turner, Clancy, Mcquade, & Cardenas, 1990, $n \leq 25$ for each group), and two studies involved complex interventions for which the independent contribution of the psychological component could not be isolated. One study involved massage as the treatment, while in another study massage was the control condition.

Message to Consumers

Nonetheless, Hoffman et al. claimed positive effects for psychological interventions contrasted with various control groups, for pain intensity, pain related interference, health-related quality of life, and depression, plus evidence of efficacy of CBT and self-regulatory treatments. Claims were made for the efficacy of multidisciplinary approaches including a psychological component, at least in terms of short-term effects on pain interference and long-term effects on return to work. "The robust nature of these findings should encourage confidence among clinicians and researchers alike." (pp. 8). Given the methodological and conceptual shortcomings in the meta-analysis, such confidence is unwarranted.

Dixon et al. (2007)

Overview

Dixon et al. reported reviewing 27 RCTs with 33 active intervention groups testing the effects of psychosocial interventions on pain in adult patients diagnosed with arthritis. They reported an overall standardized effect size of 0.18 for pain reduction.

Transparency of Reporting

In general, a lack of transparency made it difficult to assess independently the results presented by Dixon et al. First, only 15 studies (with 20 intervention groups) of the 27 studies were actually included in the meta-analysis. The authors did not explain the loss of 12 studies with 13 active intervention groups that they described as "represented in the final analysis" and for whom demographic data were included in summary descriptive statistics, but it appears from appendixes that 10 studies with 11 intervention groups were excluded because they did not report basic data for postintervention pain outcomes. It is noteworthy that only one of the 11 discarded intervention groups obtained significant results for pain. Many of the studies they excluded, apparently for the lack of published means and standard deviations, had, in fact, been included in a previous meta-analyses (e.g., Warsi, LaValley, Wang, Avorn, & Solomon, 2003).

Second, they did not present design characteristics of each study. Even in the Data Extraction tables of Appendix 2, only minimal data are presented, often discrepant with what is included in the published meta-analysis. Dixon et al. indicated that "Of the studies extracted to evidence tables, most demonstrated high quality for both internal and external validity" (p. 244). Procedures for rating study quality are described and the specific studies are rated in an appendix. We did not systematically rerate the studies. However, we did note that Barlow, Turner, and Wright (2000) received a perfect quality rating despite losing more patients from the intervention arm (77 of 311, 25%) by the 4-month follow-up than the number of patients in the intervention groups from all other studies except Lin et al. (2003). Previous systematic reviews of psychological interventions for arthritis pain (Astin, Beckner, Soeken, Hochberg, & Berman, 2002; Riemsma, Taal, Kirwan, & Rasker, 2004) have found the RCTs included by Dixon et al. to be of generally low quality and have provided specific reasons for why they were downgraded.

Third, decisions with respect to the inclusion criteria for interventions were unclear. Most of the 15 studies included in Dixon et al.'s review provided CBT or stress management led by psychologists or advanced psychology trainees. Three studies reviewed by Dixon et al. (Barlow et al., 2000; Hammond & Freeman, 2001; Riemsma, Taal, & Rasker, 2003) were described in the original studies as self-management or educational interventions. They included CBT components, but were delivered by nonpsychologists, including lay group leaders who typically had arthritis (Barlow et al., 2000) or a variety of medical staff, including nurses occupational therapists, or physical therapists (Hammond & Freeman, 2001; Riemsma et al., 2003). It is not clear why these three self-management/educational interventions were included, and others excluded. Barlow et al. (2000) and Riemsma et al. (2003) tested Lorig's Arthritis Self-Management Program. Other RCTs that have tested the same program were not identified as eligible studies by Dixon et al. (e.g., Lorig et al., 1986; Lorig et al., 1989), but no explanation was provided for this. One may also wonder how similar these three self management/educational interventions were to the control group of Keefe et al. (1996), which was described as a 10-session educational intervention on the nature and management of arthritis.

Sample Sizes of Original Studies

The largest study included in the meta-analysis was a secondary analysis of a 12-month collaborative care intervention for depressed primary care patients (Lin et al., 2003), which Dixon et al. misclassified as a psychodynamic intervention. In fact, the patients received depression care from a nurse or psychologist working with a physician. The psychotherapy intervention that was available to patients was a CBT problem-solving therapy. However, most patients in the intervention arm were using an antidepressant medication at 12 months, and less than half received specialty mental health services or psychotherapy, making it impossible to evaluate any independent contribution of the psychotherapy. It is also noteworthy that the Lin et al. study had at least nine times more patients in the intervention group than all but one of the other studies (Barlow et al., 2000) reviewed by Dixon et al. and accounted for almost half of all patients included in tabulating the summary effect size. The study should have been excluded from the meta-analysis.

Eleven studies (15 interventions) were psychological interventions delivered by trained psychologists or advanced psychology students. None of the intervention groups met the threshold cell size of $\geq$35. We found substantial baseline differences in pain for nine of 20 comparisons (e.g., Hedges's $g \geq 0.10$, including Hedges's $g > 0.75$ in 4 cases), and posttreatment differences in pain scores may have reflected the persistence of baseline differences in many cases.

Other Problems

For studies to be included in the review, patients did not have to meet a threshold criterion for pain, the primary objective of the intervention did not have to involve pain reduction, and pain did not have to be a primary outcome. Dixon et al. did not conduct a systematic review of the secondary outcomes. Rather, they reported other outcomes if data happened to be available in the same

articles as the pain outcomes. This may be the rationale for including the Lin et al. (2003) secondary analysis of collaborative care for depression. However, because depression outcomes for Lin et al. were reported in another paper, the effect was not included in Dixon et al.'s meta-analysis of depression outcomes. This suggests what could be a more pervasive problem in calculating effect sizes for outcomes other than pain due to different outcomes being reported in separate papers.

We found numerous discrepancies in reporting details of studies and erroneous calculations of effect sizes in the published meta-analysis. Not all of the studies presented in Dixon et al.'s Figure 1 are even listed in Appendix 2, the Data Extraction Table. Sample sizes for at least half a dozen of the studies listed in Figure 1 do not match the sample sizes reported in the original studies. For instance, Sharpe et al. (2001) explicitly indicate that they report data in their tables for 23 intervention and 22 control completers. Dixon et al. report sample sizes of 19 and 18. Dixon et al. do not indicate how they selected pain measures for the meta-analysis when studies provided more than one measure.

The authors do not state how effect sizes were calculated, but reanalysis suggests that post-treatment means and standard deviations were used when possible, with change scores used if post-treatment data were not provided. When the latter was apparently done for the Riemsma et al. (2003) study at 12-months, however, the group education effect was zero, but listed as −0.17 in Dixon et al.'s Figure 1. The effect size listed by Dixon et al. for the group education with spousal involvement from the Riemsma study was correct, but in the wrong direction: the intervention group had worse pain at the outcome. Overall, there were five intervention groups erroneously listed in Figure 1 by Dixon et al. with negative effect sizes (lower pain in intervention group) when the intervention groups had higher pain levels (CBT2 from Keefe et al., 2004; Kraaimaat, Brons, Geenen, & Bijlsma, 1995; CBT1 from Radojevic, Nicassio, & Weisman, 1992; Shearn & Fireman, 1985; CBT1 Riemsma et al., 2003).

Message to Consumers

Dixon et al. claimed, "These findings indicate that psychosocial interventions may have significant effects on pain and other outcomes in arthritis patients. Ample evidence for the additional benefit of such interventions over and above that of standard medical care was found" (p. 241). They also recommended, ". . . it is important that arthritis patients be made aware that although psychosocial interventions appear to have some effects on pain, these treatments are most likely to enhance their quality of life by producing improvements in other important areas such as coping, anxiety, pain, self-efficacy, depression, joint swelling, and physical disability" (p. 248). A Clinician Commentary (Pisetsky, 2007) accompanying the article echoed this assessment: "As described in the article . . . well-designed and well-controlled trials have conclusively established the value of various psychosocial interventions . . . in reducing arthritis pain. Although the literature on cognitive—behavioral therapy is the most extensive, data are available to support the utility of all modalities tested" (p. 657). Confidence in these conclusions, however, is mitigated by the methodological issues and problems noted above. Even if one accepted the basic results of the meta-analysis, one would not know what to do with them. Should psychological services be provided by doctoral level providers? Or will self-management programs run by nurses do about as well? The review by Dixon et al. does not answer these questions nor clarify to which type of intervention their results apply.

Jacobsen et al. (2007)

Overview

Jacobsen et al. (2007) conducted their meta-analysis to determine the efficacy of psychological and activity-based interventions for cancer-related fatigue. They reported a statistically significant overall effect size of $d = 0.09$, as well as a statistically significant effect for psychological interventions ($d = 0.10$), and a nonsignificant effect for activity-based interventions ($d = 0.05$). Jacobsen and colleagues reviewed 24 psychological and 17 activity-based trials for their meta-analysis of interventions for cancer-related fatigue. Nineteen studies provided sufficient information for inclusion in the meta-analysis and another 11 could be included after additional information was obtained from the authors of the original studies. In total, 18 psychological and 12 activity-based studies were included in the meta-analysis.

Previous Meta-Analyses

Kangas, Bovbjerg and Montgomery (2008) claimed that Jacobsen et al. failed to identify over 40 additional trials meeting their criteria, which might have been due to a restricted search strategy including the use of only PsycINFO, MEDLINE, and CINAHL. While it is beyond the purposes of this article to resolve this important discrepancy, we did find that 20 RCTs published before Jacobsen et al. completed their search were left out by Jacobsen et al., but included in Kangas et al. (2008).

Previous smaller meta-analyses (2 to 11 trials) of nonpharmacological interventions for cancer-related fatigue which included non-RCTs, reported slightly higher effect sizes between 0.11 and 0.24 (Conn, Hafdahl, Porock, McDaniel, & Nielsen, 2006; Luebbert, Dahme, & Hasenbring, 2001; Schmitz et al., 2005). The recent meta-analysis by Kangas et al. (2008) reported a much higher effect size of 0.34 based on 57 RCTs. No significant difference was found by Kangas et al. for psychological versus activity-based interventions. Besides the difference in number of studies included, another reason for the difference in overall effect size may be that Jacobsen et al. (2007) utilized the final follow-up measurement while Kangas et al. (2008) utilized the first follow-up measurement. However, discrepancies in effect size were also present for activity-based interventions that mostly reported outcomes immediately after the intervention (0.05 in the Jacobsen et al. study vs. 0.42 for fatigue and 0.69 for vigor in the Kangas et al. study). The conflicting results of these two meta-analyses highlight the subjectivity of meta-analysis in terms of search strategy, inclusion criteria and integration; different decisions may lead to very different conclusions.

Methodological Quality of Included Studies

A supplementary table of methodological ratings published online revealed that the quality of 70%–80% of the studies included in the Jacobsen et al. meta-analysis was rated as only fair. Only

40% of the studies included in the meta-analysis presented statistics showing that the intervention and control groups were comparable at baseline; fewer than half indicated the number of patients enrolled, treatment drop-out and reasons for drop-out; and only 4% of the psychological and 29% of the activity-based intervention trials provided intent-to-treat analyses.

Sample Sizes of Original Studies

Appendixes with information on the original studies showed that six of the 18 psychological and seven of the 12 activity-based intervention studies included in Jacobsen et al.'s meta-analysis fell below $n \geq 35$. Of the activity-based intervention studies, four studies had a cell size of 13 or less. Of the 12 psychological interventions with at least 35 patients per cell, Bordeleau et al. (2003) and Goodwin et al. (2001) reported on the same sample.

Clinical Heterogeneity

The interventions evaluated in the included RCTs were quite diverse, including supportive expressive group therapy, mindfulness meditation in a group setting or group psychoeducation; group or individual CBT; telephone education about energy conservation and activity management; individual pain education and management; individual cancer health education; individual stress management, or the provision of an audiotape of a consult with the oncologist. Such clinical heterogeneity precludes a meaningful answer as to the intervention type to which an effect can be generalized.

Of the five activity-based intervention studies with cell sizes of at least 35, one study compared a group who received psychotherapy to a group receiving psychotherapy plus exercise. Another study reported a larger decrease in fatigue in the control than the exercise condition. Two studies focused on supervised exercise, while the other three studies focused on home-based exercise. Thus, clinical heterogeneity noted in the intervention conditions extended to the comparison-control conditions as well.

Other Problems

The most serious criticism of Jacobsen et al. is that few of the RCTs included in the meta-analysis had the specific aim of reducing fatigue or even stated fatigue reduction as a relevant hypothesis. At best, such interventions might be construed as active comparison-control treatments and contrasted, rather than integrated with treatments explicitly targeting fatigue. In only three of the 18 (17%) psychological intervention studies was fatigue a primary outcome and in only five of the 12 (42%) activity-based intervention studies. In no studies was there a minimal threshold of fatigue as an entry criterion. Kangas et al. (2008) is subject to the same criticism, but distinguished between RCTs that had aims or hypotheses concerning fatigue and those that did not. In their larger sample of studies, only 28% had a specific fatigue-related aim or hypothesis, and of these, half had positive effects, a substantially greater proportion than RCTs without such an aim or hypothesis.

Many of the RCTs that did not have aims or hypotheses related to fatigue were selected because they included the Profile of Mood States (POMS) among their outcomes, which has subscales assessing fatigue and vigor-activity. Studies targeting distress commonly include the POMS as part of a battery of distress measures. Fatigue and vigor-activity are often included as an outcome for RCTs without an explicit hypothesis simply because these subscales are part of the instrument. Outcomes in these studies are reported without a prioritizing (primary vs. secondary) of these multiple measures and with a strong confirmatory bias (See Coyne et al., 2006). Whether results for fatigue and vigor-activity subscales are reported at all may depend on whether a case can otherwise be made for the efficacy of the intervention for distress in terms of depression-dejection or anxiety-tension subscales of the POMS. Thus, results for fatigue and vigor-activity may only be selectively available. Overall, however, we find that Jacobsen et al.'s decisions to include RCTs simply on the basis of a relevant published outcome measure, not on the basis of the aim or hypothesis of the study and their failure to distinguish aims or hypotheses of studies in conducting moderator analyses undercuts any claims of clinical or pragmatic implications of the results.

Message to Consumers

Jacobsen and colleagues claimed: "results of this review provide limited support for the clinical use of nonpharmacological interventions to prevent or relieve cancer-related fatigue" and ". . . evidence of efficacy is stronger for psychological interventions than for activity-based interventions." (p. 665). We did not find any evidence in the article, however, to suggest that there is a difference in efficacy of psychological versus activity-based interventions. Specifically, the moderator analysis for type of intervention was not significant. In fact, none of the moderator analyses were significant, including comparisons of the efficacy based on whether fatigue or vigor was measured as an outcome, whether participants had been diagnosed with breast cancer or another type of cancer, whether interventions were delivered to individuals or a group, and whether interventions were home-based or supervised. The overall effect size that Jacobsen et al. reported, 0.09 (95% CI = .02–.16) which translates to an r^2 of .002, indicates a negligible effect. If accepted at face, this result should discourage offering psychosocial services to patients with cancer for relief of fatigue. It is not clear from the meta-analysis provided by Jacobsen et al., however, if this would be the correct decision.

Integration and Commentary

Consistent with our experience with the meta-analysis by Zimmermann and colleagues (2007), evaluating these four meta-analyses of interventions for adults (Dixon et al., 2007; Hoffman et al., 2007; Irwin et al., 2006; and Jacobsen et al., 2007) that appeared in a new section of *Health Psychology*, Evidence Based Treatment Reviews, was an arduous, labor intensive and ultimately frustrating process. We started with the published meta-analyses and found numerous lapses in transparency and notable inconsistencies, miscalculations, and contradictions. We reviewed additional sources, such as the original RCTs, but also related studies and meta-analyses often covering the same literatures. Our sense increased that there were fatal flaws in these meta-analyses related to inaccuracies, failure to adequately address issues of quality and arbitrariness of study selection and inclusion. We doubt whether a casual reader or a consumer in need of quick clinically or policy

relevant summaries would be motivated or alert to the need to undertake such a reappraisal. In all likelihood, they would defer to the evaluations of the literature usually confidently offered by meta-analyses and discouraged by the opaqueness with regard to key methods and decisions from questioning further.

Many of the problems encountered in these meta-analyses stemmed from the notable inadequacies in the conduct and reporting of the original RCTs considered for inclusion. Quality was often only poor to fair at best, and these studies also tended to be underpowered, compounding the problems by inadequacies in the design, analyses, and reporting of these studies. While these meta-analyses sometimes acknowledged the problem with the original literature and quantified the problem with formal ratings, no effort was made in any analysis to exclude poor quality studies or to perform sensitivity analyses evaluating effects of whether they were excluded. We found evidence of studies identified in search procedures as eligible for inclusion not being included in the meta-analysis, some of the excluded studies nonetheless having been included in other published meta-analyses (Dixon et al., 2007; Jacobsen et al., 2007), and of results being more accessible in original studies when they were favorable to the evaluation of an intervention (e.g., Dixon et al.). Thus, a confirmatory bias in the original studies was compounded by a confirmatory bias in whether they were included in these meta-analyses. We had to turn to the original literature before identifying some problems. For instance, none of the RCTs evaluating multi-component interventions for back pain allowed isolation of the independent contribution of the psychological component. In another case, by far the largest RCT included in Dixon et al. study (2007), which accounted for almost 50% of patients whose data were meta-analyzed, similarly did not allow isolation of the contribution of psychotherapy. All of these problems speak to the subjectivity of meta-analysis and high likelihood that meta-analyses of behavioral medicine interventions at this time would benefit from more careful attention to methodological issues, including when it is appropriate to conduct a meta-analysis, given the confirmatory bias and limitations of the existing trials, most of which are not of high quality.

We found support for our proposal that small studies with less than 55% power to detect an effect even when it was present should be excluded. First, some analyses heavily depended on studies so small that there was substantially less than 55% probability of moderate effects being identified. Second, there was evidence that the poor quality of the small RCTs influenced whether effects were available for scrutiny. For instance, some meta-analyses were influenced by pre-post differences that were more accurately attributable to the intervention and control groups not being equivalent at baseline, and in at least one instance, a substantial number of small RCTs were not entered into the meta-analysis because sufficient data for meta-analysis from these nonsignificant trials were not made available in the original articles. While serious methodological problems should be of concern for any sized study, they can become more decisive in their impact on small studies and whether the results become available in published articles. While we can readily detect that such biases exist, our inability to specify the extent of bias precludes any confidence in efforts to compensate for bias statistically (Kraemer et al., 1998).

Do these meta-analyses supply a suitable foundation for clinical and policy decisions? Do they indicate that psychosocial interventions should be staffed and funded by third party payments? Or which ones? For numerous reasons, we do not believe the meta-analyses we reviewed provide a basis for evaluating the potential efficacy of such interventions. At best, it would be premature to make decisions on the basis of their results. The first set of issues lies in a lack of transparency and other substantive and methodological problems with which the analyses were conducted and reported, including their being conducted with unrepresentative subsamples of the available literature. Second, the meta-analyses depend on small numbers of underpowered, methodologically inadequate trials, particularly for any moderator analyses comparing alternative treatments. Third, the reviews do not establish standards for clinically meaningful effects. Jacobsen et al. (2007); Hoffman et al. (2007), and Dixon et al. (2006) claim what would reasonably be considered clinically negligible effects without identifying them as such.

Our review of these meta-analyses has a number of notable limitations. It was important that we did not undertake a systematic re-review of the relevant literatures, nor recalculate every effect size we encountered. Often, our probing of the literature was initiated because of a suspicion about what was presented in the published meta-analyses and available supplementary materials. Also, it is likely that with extended dialogue with the authors, we could have clarified additional points and even resolved some ambiguities and apparent contradictions in the published papers. However, our goal was not to characterize the literatures covered by these meta-analyses, but to conduct an appraisal of the adequacy and accuracy of these meta-analyses in summarizing these literatures. In doing so, we were more concerned with uncovering any evidence that the meta-analyses were inadequate, undependable or likely biased as the foundations for practical decision-making. Our view is that meta-analyses should be held to high standards. As a secondary literature, there is the likelihood that their conclusions will be accepted in lieu of a scrutiny of the primary RCTs, particularly when these conclusions are presented confidently and persuasively and without adequate transparency. There is the potential that particular claims based on the meta-analyses about the relative efficacy of interventions will be used to override patient and clinician preferences and restrict the availability of interventions, regardless of their accuracy of these claims. Such claims can also influence research priorities by indicating that particular empirical questions have been settled and so no further funding should be allocated to addressing them.

We think that results of our exercise should serve as a wake up call: prospective authors of meta-analyses, reviewers, and anticipated consumers should be educated about how to conduct and report a meta-analysis and the numerous threats to their validity. We do not know how generalizable our criticisms are to other meta-analyses of behavioral medicine interventions, but they do provide a basis for heightened skepticism. Overall, we would like to see more attention in the behavioral medicine literature to the controversies and developing standards for doing and reporting meta-analyses that are appearing in clinical epidemiological and biomedical journals. However, the behavioral medicine community should be active participants in the process of improving the process of evaluating interventions, not just spectators, and its key journals need to provide a forum for debate tailored to the specific needs and interests of the field. Recently, elaborate new American Psychological Association standards were announced for reporting both clinical trials and systematic reviews and meta-analyses

(Cooper, Maxwell, Stone, & Sher, 2008). Even before these standards are adopted as requirements for publication in APA journals and elsewhere, they provide invaluable guidance for avoiding some of the problems we identified with these four meta-analyses. Yet, it will be decades before these new standards are seen in the quality of the bulk of the trials available for consideration for inclusion in meta-analyses. Meanwhile, authors contemplating conducting a meta-analysis must contend with a literature dominated by methodologically flawed studies and decide whether and how to proceed. One way to resolve this dilemma is not to conduct a meta-analysis, as Moher, Jadad, and Klassen (1998) suggest:

> There is often a perception that the statistical combination of data across studies is the most important part of a systematic review. We take such a view cautiously. We believe that a well-reported, systematic qualitative review is much better than an inappropriately conducted and reported quantitative review or meta-analysis (pp. 916).

The risk of publishing meta-analyses that are premature because of limitations in the number and quality of available studies is that the accumulation of better, larger-scale studies and the integration of their results in a future meta analysis will be discouraged because of a false impression that the research question was already settled.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Alaranta, H., Rytokoski, U., Rissanen, A., Talo, S., Ronnemaa, T., Puukka, P., et al. (1994). Intensive physical and psychosocial training-program for patients with chronic low-back-pain—a controlled clinical-trial. *Spine, 19,* 1339–1349.

Astin, J. A., Beckner, W., Soeken, K., Hochberg, M. C., & Berman, B. (2002). Psychological interventions for rheumatoid arthritis: A meta-analysis of randomized controlled trials. *Arthritis & Rheumatism-Arthritis Care & Research, 47,* 291–302.

*Barlow, J. H., Turner, A. P., & Wright, C. C. (2000). A randomized controlled study of the Arthritis Self-Management Programme in the UK. *Health Education Research, 15,* 665–680.

*Bordeleau, L., Szalai, J. P., Ennis, M., Leszcz, M., Speca, M., Sela, R., et al. (2003). Quality of life in a randomized trial of group psychosocial support in metastatic breast cancer: Overall effects of the intervention and an exploration of missing data. *Journal of Clinical Oncology, 21,* 1944–1951.

*Brox, J. I., Sorensen, R., Friis, A., Nygaard, O., Indahl, A., Keller, A., et al. (2003). Randomized clinical trial of lumbar instrumented fusion and cognitive intervention and exercises in patients with chronic low back pain and disc degeneration. *Spine, 28,* 1913–1921.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Conn, V. S., Hafdahl, A. R., Porock, D. C., McDaniel, R., & Nielsen, P. J. (2006). A meta-analysis of exercise interventions among people treated for cancer. *Supportive Care in Cancer, 14,* 699–712.

Cooper, H., Maxwell, S., Stone, A., & Sher, K. J. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851.

Coyne, J. C., Lepore, S. J., & Palmer, S. C. (2006). Efficacy of psychosocial interventions in cancer care: Evidence is weaker than it first looks. *Annals of Behavioral Medicine, 32,* 104–110.

Coyne, J. C., Thombs, B., & Hagedoorn, M. (2008). A meta-analysis of psychosocial interventions for cancer patients gone awry. *Annals of Behavioral Medicine, 37,* 94–96.

Dixon, K. E., Keefe, F. J., Scipio, C. D., Perri, L. M., & Abernethy, A. P. (2007). Psychological interventions for arthritis pain management in adults: A meta-analysis. *Health Psychology, 26,* 241–250.

*Goodwin, P. J., Leszcz, M., Ennis, M., Koopmans, J., Vincent, L., Guther, H., et al. (2001). The effect of group psychosocial support on survival in metastatic breast cancer. *New England Journal of Medicine, 345,* 1719–1726.

*Hammond, A., & Freeman, K. (2001). One-year outcomes of a randomized controlled trial of an educational-behavioral joint protection program for people with rheumatoid arthritis. *Rheumatology, 40,* 1044–1051.

*Hernandez-Reif, M., Field, T., Krasnegor, J., & Theakston, H. (2001). Lower back pain is reduced and range of motion increased after massage therapy. *International Journal of Neuroscience, 106,* 131–145.

Higgins, J. P. T., & Altman, D. G. (2008). chap. 8: Assessing risk of bias in included studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* version 5.0.1 (updated September 2008). The Cochrane Collaboration. Available from www.cochrane-handbook.org

Hoffman, B. M., Papas, R. K., Chatkoff, D. K., & Kerns, R. D. (2007). Meta-analysis of psychological interventions for chronic low back pain. *Health Psychology, 26,* 1–9.

Howard, G. S., Hill, T. L., Maxwell, S. E., Baptista, T. M., Farias, M. H., Coelho, C., et al. What's wrong with research literatures? And how to make them right. *Review of General Psychology, 13,* 146–166.

Ioannidis, J. P. (2008a). Why most discovered true associations are inflated. *Epidemiology, 19,* 640–648.

Ioannidis, J. P. A. (2008b). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice, 14,* 951–957.

Irwin, M. R., Cole, J. C., & Nicassio, P. M. (2006). Comparative meta-analysis of behavioral interventions for insomnia and their efficacy in middle-aged adults and in older adults 55+ years of age. *Health Psychology, 25,* 3–14.

Jacobsen, P. B., Donovan, K. A., Vadaparampil, S. T., & Small, B. J. (2007). Systematic review and meta-analysis of psychological and activity-based interventions for cancer-related fatigue. *Health Psychology, 26,* 660–667.

Juni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care—Assessing the quality of controlled clinical trials. *British Medical Journal, 323,* 42–46.

Juni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association-Journal of the American Medical Association, 282,* 1054–1060.

Kangas, M., Bovbjerg, D. H., & Montgomery, G. H. (2008). Cancer-related fatigue: A systematic and meta-analytic review of non-pharmacological therapies for cancer patients. *Psychological Bulletin, 134,* 700–741.

*Kankaanpaa, M., Taimela, S., Airaksinen, O., & Hanninen, O. (1999). The efficacy of active rehabilitation in chronic low back pain—Effect on pain intensity, self-experienced disability, and lumbar fatigability. *Spine, 24,* 1034–1042.

Keefe, F. J., Caldwell, D. S., Baucom, D., Salley, A., Robinson, E., Timmons, K., et al. (1996). Spouse-assisted coping skills training in the management of osteoarthritic knee pain. *Arthritis Care and Research, 9,* 279–291.

*Keefe, F. J., Blumenthal, J., Baucom, D., Affleck, G., Waugh, R., Caldwell, D. S., et al. (2004). Effects of spouse-assisted coping skills training and exercise training in patients with osteoarthritic knee pain: A randomized controlled study. *Pain, 110,* 539–549.

*Kraaimaat, F. W., Brons, M. R., Geenen, R., & Bijlsma, J. W. J. (1995). The effect of cognitive-behavior therapy in patients with rheumatoid-arthritis. *Behavior Research and Therapy, 33,* 487–495.

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis:

Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

*Lichstein, K. L., Riedel, B. W., Wilson, N. M., Lester, K. W., & Aguillard, R. N. (2001). Relaxation and sleep compression for late-life insomnia: A placebo-controlled trial. *Journal of Consulting and Clinical Psychology, 69,* 227–239.

*Lin, E. H. B., Katon, W., Von Korff, M., Tang, L. Q., Williams, J. W., Kroenke, K., et al. (2003). Effect of improving depression care on pain and functional outcomes among older adults with arthritis—A randomized controlled trial. *Journal of the American Medical Association-Journal of the American Medical Association, 290,* 2428–2434.

*Lorig, K., Lubeck, D., Kraines, R. G., Seleznick, M., & Holman, H. R. (1985). Outcomes of self-help education for patients with arthritis. *Arthritis & Rheumatism*, *28*, 680–685.

*Lorig, K., Seleznick, M., Lubeck, D., Ung, E., Chastain, R. L., & Holman, H. R. (1989). The beneficial outcomes of the Arthritis Self-Management Course are not adequately explained by behavior change. *Arthritis & Rheumatism, 32*, 91–95.

Luebbert, K., Dahme, B., & Hasenbring, M. (2001). The effectiveness of relaxation training in reducing treatment-related symptoms and improving emotional adjustment in acute non-surgical cancer treatment: A meta-analytical review. *Psycho-Oncology, 10,* 490–502.

*Manne, S. L., Girasek, D., & Ambrosino, J. (1994). An evaluation of the impact of a cosmetics class on breast cancer patients. *Journal of Psychosocial Oncology, 12,* 83–99.

*McQuellon, R. P., Wells, M., Hoffman, S., Craven, B., Russell, G., Cruz, J., et al. (1998). Reducing distress in cancer patients with an orientation program. *Psycho-Oncology, 7,* 207–217.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *Lancet, 354,* 1896–1900.

Moher, D., Jadad, A. R., & Klassen, T. P. (1998). Guides for reading and interpreting systematic reviews: III How did the authors synthesize the data and make their conclusions? *Archives of Pediatrics and Adolescent Medicine, 152,* 915–920.

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet, 357,* 1191–1194.

Newell, S. A., Sanson-Fisher, R. W., & Savolainen, N. J. (2002). Systematic review of psychological therapies for cancer patients: Overview and recommendations for future research. *Journal of the National Cancer Institute, 94,* 558–584.

*Nezu, A. M., Nezu, C. M., Felgoise, S. H., McClure, K. S., & Houts, P. S. (2003). Project genesis: Assessing the efficacy of problem-solving therapy for distressed adult cancer patients. *Journal of Consulting and Clinical Psychology, 71,* 1036–1048.

O'Connor, D., Green, S., & Higgins, J. P. T. (2008). chap. 5: Defining the review question and developing criteria for including studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane Handbook of Systematic Reviews of Intervention* version 5.0.1 (updated September 2008). The Cochrane Collaboration. Available from www.cochrane-handbook.org

Pisetsky, D. S. (2007). Clinician's comment on the management of pain in arthritis. *Health Psychology, 26,* 657–659.

*Radojevic, V., Nicassio, P. M., & Weisman, M. H. (1992). Behavioral intervention with and without family support for rheumatoid-arthritis. *Behavior Therapy, 23,* 13–30.

*Riemsma, R. P., Taal, E., & Rasker, J. J. (2003). Group education for patients with rheumatoid arthritis and their partners. *Arthritis & Rheumatism-Arthritis Care & Research, 49,* 556–566.

Riemsma, R. P., Taal, E., Kirwan, J. R., & Rasker, J. J. (2004). Systematic review of rheumatoid arthritis patient education. *Arthritis & Rheumatism-Arthritis Care & Research, 51,* 1045–1059.

*Rybarczyk, B., Lopez, M., Benson, R., Alsten, C., & Stepanski, E. (2002). Efficacy of two behavioral treatment programs for comorbid geriatric insomnia. *Psychology and Aging, 17,* 288–298.

Sackett, D. L., & Cook, D. J. (1993). Can we learn anything from small trials. *Doing More Good Than Harm: The Evaluation of Health Care Interventions, 703,* 25–32.

Schmitz, K. H., Holtzman, J., Courneya, K. S., Masse, L. C., Duval, S., & Kane, R. (2005). Controlled physical activity trials in cancer survivors: A systematic review and meta-analysis. *Cancer Epidemiology Biomarkers & Prevention, 14,* 1588–1595.

Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *Lancet, 365,* 1348–1353.

*Sharpe, L., Sensky, T., Timberlake, N., Ryan, B., Brewin, C. R., & Allard, S. (2001). A blind, randomized, controlled trial of cognitive-behavioral intervention for patients with recent onset rheumatoid arthritis: Preventing psychological and physical morbidity. *Pain, 89,* 275–283.

Sheard, T., & Maguire, P. (1999). The effect of psychological interventions on anxiety and depression in cancer patients: Results of two meta analyses. *British Journal of Cancer, 80,* 1770–1780.

*Shearn, M. A., & Fireman, B. H. (1985). Stress management and mutual support groups in rheumatoid-arthritis. *American Journal of Medicine, 78,* 771–775.

Staines, G. L., & Cleland, C. M. (2007). Bias in meta-analytic estimates of the absolute efficacy of psychotherapy. *Review of General Psychology, 11,* 329–347.

*Turner, J. A., Clancy, S., Mcquade, K. J., & Cardenas, D. D. (1990). Effectiveness of behavioral-therapy for chronic low-back-pain—A component analysis. *Journal of Consulting and Clinical Psychology, 58,* 573–579.

Wanous, J. P., Sullivan, S. E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74,* 259–264.

Warsi, A., LaValley, M. P., Wang, P. S., Avorn, J., & Solomon, D. H. (2003). Arthritis self-management education programs—A meta-analysis of the effect on pain and disability. *Arthritis and Rheumatism, 48,* 2207–2213.

Zimmermann, T., Heinrichs, N., & Baucom, D. H. (2007). "Does one size fit all?" Moderators in psychosocial interventions for breast cancer patients: A meta-analysis. *Annals of Behavioral Medicine, 34,* 225–239.